PERGAMON

Information Processing and Management 35 (1999) 317–336

INFORMATION PROCESSING & MANAGEMENT

Content locality in distributed digital libraries

Charles L. Viles[a,*], James C. French[b]

[a]School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA
[b]Department of Computer Science, University of Virginia, Charlottesville, VA 22903, USA

Abstract

In this paper we introduce the notion of content locality in distributed document collections. Content locality is the degree to which content-similar documents are colocated in a distributed collection. We propose two metrics for measurement of content locality, one based on topic signatures and the other based on collection statistics. We provide derivations and analysis of both metrics and use them to measure the content locality in two kinds of document collections, the well-known TREC corpus and the Networked Computer Science Technical Report Library (NCSTRL), an operational digital library. We also show that content locality can be thought of temporally as well as spatially and provide evidence of its existence in temporally ordered document collections like news feeds. © 1999 Elsevier Science Ltd. All rights reserved.

1. Introduction

Successful design, testing and deployment of digital libraries involves research in a variety of disciplines, including information retrieval, databases, collection development, archival policies, human computer interaction, intellectual property and commerce models to name just a few. Here we consider the digital library (DL) as a set of autonomous, distinct document collections that 'cooperate' to support search and retrieval. In this distributed setting, we expect that the topical distribution of content among collections (sites) in the system will be non-uniform. For example, in a DL of the works of contemporary literature of the American South, we would expect that the materials of William Styron would reside in large part at Duke University, his alma mater, rather than be distributed uniformly throughout all member collections in the DL.

* Corresponding author. Tel.: +1-919-962-8366.
E-mail address: viles@ils.unc.edu (C.L. Viles)

0306-4573/99/$ - see front matter © 1999 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(98)00064-8

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **1999**	2. REPORT TYPE	3. DATES COVERED **00-00-1999 to 00-00-1999**

4. TITLE AND SUBTITLE **Content locality in distributed digital libraries**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Virginia,Department of Computer Science,151 Engineer's Way,Cahrlottesville,VA,22094-4740**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**20**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

A major finding in previous work in distributed information retrieval (Viles & French, 1995; French & Viles, 1996; Viles, 1996) is that such content-based allocation of documents to sites affects the quality of retrieval. When there is no inter-site communication, distributed document collections whose content is heavily skewed exhibit much poorer retrieval effectiveness than collections whose content is uniformly distributed.

We refer to this phenomenon, previously called 'content skew' (Viles, 1996), as *content-locality*. Intuitively, content-locality is the degree to which topically similar documents are co-located in a distributed document collection. In this paper, we

- provide two methods for measuring content-locality, one topically based and the other statistically based.
- measure the locality of the multi-year, multi-source TREC collection using the topically based metric.
- measure the locality of an operational distributed digital library, the Networked Computer Science Technical Report Library (NCSTRL, http://www.ncstrl.org/) (Davis, 1995) using the statistically based metric.
- show that there is also a temporal analogue to the spatial content-locality documented here.

The primary goal of this paper is to describe the nature of content-locality and quantify its presence in distributed document collections. To do so, we define methods to measure content-locality and measure the locality of two distributed document collections, one constructed for IR experimentation and one that is an operational digital library. We provide a critical analysis of each of the proposed methods with the goal of gaining additional insight into the underlying phenomenon.

The first measure we give here is topic-centric. Essentially, we treat a document collection as a set of topics and determine how each topic is allocated to the member sites of the distributed collection. The more asymmetric this allocation, the higher the locality and the more uniform, the lower the locality. The second measure is statistically based. For each site in the distributed collection, we look at the distribution of terms in the local collection and compare it against what they would be in a centralized collection composed of the contents of all local collections.

The notion of content-locality in distributed document collections is new. Each of the two possible metrics we propose here has its advantages. The topically based metric gives insight into the distribution of content *by topic* in the collection. However, the fidelity of the measurement depends upon the accuracy of topic determination. The statistically based metric is simple to calculate and requires no subjective determination of topic, thus it holds promise for use in operational systems. For both metrics there is an interpretation and scaling issue, e.g. suppose locality is measured as 0.186, does that represent a skewed or unskewed system? These problems can only be overcome once distributed document archive systems have been deployed and analyzed in realistic environments.

2. Why measure content-locality?

If we can determine that the content-locality of a distributed collection is low, then the implication from an engineering perspective is that no inter-site communication is needed in order to attain good search effectiveness (Viles & French, 1995; French & Viles, 1996). This is desirable, because it implies that a site can operate more or less independently and allows much more flexibility in the particulars of things like index structure (e.g. whether or not collection statistics are pre-computed and stored in the index) and communication infrastructure (e.g. whether or not a site must handle collection statistic updates originating 'off-site'). Conversely, if the content-locality of a collection is high then some kind of intersite communication is needed to achieve search quality commensurate with a centralized system. Engineering the system in this situation then involves an informed trade-off between the 'best' search quality and a simpler, more efficient system. On the other hand, it is also possible to exploit highly content-localized systems by quickly eliminating or deferring search at sites with low topical relevance.

The determination of content-locality itself requires inter-site communication, but it is of a different kind than what would be used operationally when sites exchanged collection statistics like document frequency or document lengths. In the latter case, the communication may require the expensive update of disk-based structures at each recipient and each site could potentially be receiving from every other site. In the former case, each site need send only it's topical or statistical descriptions to some process or entity that has the ability to integrate them into a single unified description. The availability of such descriptions is commonly assumed in much of the literature on collection selection (also called 'database selection') (Callan, Lu & Croft, 1995; Gravano, Chang, Paepcke & Garcia-Molina, 1997; Gravano & Garcia-Molina, 1997).

As has been mentioned, the topical locality measurement requires a set of topics that, taken together, define the content of the distributed collection. The set of topics itself is useful in a variety of ways, so in some respect it is equally appropriate to think of locality as one of many reasons to undertake the topical determination of a document collection. Once the topical content is determined, locality is simple to calculate.

In addition to the determination of content-locality, there are many potential benefits of knowing the topical content of a collection. These include:

Topic	Text Excerpt
161	*Document will provide information on the problems and actions associated with what is known as acid rain.*
152	*Accusations of Cheating by Contractors on U.S. Defense Projects. Document will refer to an alleged illegality committed by any entity seeking a contract on behalf of the U.S. Military Forces.*
37	*Document identifies software products which adhere to IBM's SAA standards. To be relevant, a document must identify a piece of software which is considered a Systems Application Architectural (SAA) component or one which conforms to SAA.*

Fig. 1. Three topics from the TREC collection that exhibit different degrees of topical locality. Topics are arranged from 'low' (topic 161) to 'high' (topic 37) locality.

Table 1
Summary of notation used in content-locality derivation

$b_{s,t}$	$n_{s,t}/n_t$, proportional size of topic t
c_s	N_s/N, proportional size of site s
n_t	size of topic t
$n_{s,t}$	size of topic t at site s
N	collection size
N_s	collection size at site s
S	number of sites
$\mathscr{T}$	set of topics
T	number of topics
δ_s	statistical locality for site s
δ	statistical locality for system
$\sigma_{s,t}$	locality for topic t at site s
σ_t	locality for topic t
σ	topical locality for system

- Topic tracking: looking for topics that are new, 'hot', or have very little or very much activity.
- Efficiency: caching documents in the same topic together or nearby.
- Intelligent browsing: in interactive systems, a topical map can provide a set of related starting points for browsing (Cutting, Karger, Pedersen, & Tukey, 1992; Kellogg & Subhas, 1996).
- Pre-fetching: the fetching of one or more documents in the same topic might signal the system to start pre-fetching other documents in the topic.

3. Topic-based locality

Before presenting the details of the topical locality measure, we present three topics from the TREC experiments (Harman, 1995) in Fig. 1. When considered in the context of the heterogeneous mix of sources that comprise the TREC corpus, these topics provide further intuition into the nature of locality. Later (Section 7), we provide specific locality measurements for 250 of the TREC topics. In the case of this example, the topics exhibited 'low' (topic 161), 'average' (topic 152) and 'high' (topic 37) topical locality.

The notation we use in this paper is given in Table 1.

The locality measure we define in this section is topic-centered. To determine locality for the entire system, we first determine locality for each individual topic and then combine these to get a system level measurement of content-locality.

3.1. Individual topic locality

Intuitively, there are two contributing factors to topic-based content locality. First, given some topic t, the total number of sites k that contain some member of t affect locality. If k is small, then locality should be high, if k is large, then locality should be low. Second, given t is represented at k sites, the more asymmetric the distribution of members of t at the k sites, the more content-localized the system is.

The general approach we take is to define content-locality for a topic as the sum of squared error terms where 'error' is the distance from a content-uniform system, one where a topic is equally represented at all sites.

The locality for topic t at site s is denoted by $\sigma_{s,t}$ and is calculated by

$$\sigma_{s,t} = (b_{s,t} - E[b_{s,t}])^2, \qquad (1)$$

where $b_{s,t} = n_{s,t}/n_t$ is the size $n_{s,t}$ of the topic at the site relative to the overall size n_t of that topic. Here we think of $E[b_{s,t}]$ as the expected value of $b_{s,t}$ when content is uniformly distributed throughout the distributed collection. In content-uniform collections, we would expect that $E[b_{s,t}]$ would track the proportionate size of the collection at site s. If $c_s = N_s/N$ is the proportionate size, then

$$\forall t,\ E[b_{s,t}] = c_s$$

so by substitution Eq. (1) becomes

$$\sigma_{s,t} = (b_{s,t} - c_s)^2$$

Locality for some topic t is denoted σ_t and is determined by summing the locality for that topic at each site and taking the square root. So

$$\sigma_t = \sqrt{\sum_{s=1}^{S} \sigma_{s,t}}$$

and by substitution

$$\sigma_t = \sqrt{\sum_{s=1}^{S} (b_{s,t} - c_s)^2}.$$

In Appendix 1, we show that for any distributed collection, $0 \leq \sigma_t < \sqrt{2}$, and for collections where each site is approximately the same size, $0 \leq \sigma_t \leq 1$.

3.1.1. Behavior of σ_t

We would like the measure of content-locality to reflect the two contributing factors to locality that we outlined previously, namely:

1. As fewer (more) *sites* contain members of some topic t, measured locality should increase (decrease).
2. Given that k sites contain members of t, the more asymmetric the distribution of these members, the higher measured locality should be.

In Appendix 1 we show that property (1) is followed. Specifically, if we consider the system where a topic is evenly distributed over k sites, then if all sites have about the same number of documents

$$\sigma_t = \sqrt{1/k - 1/S}.$$

As $k \rightarrow S$, $\sigma_t \rightarrow 0$.

Now consider property (2). The measured locality for a topic that has members at k sites should become higher as the distribution becomes more non-uniform or asymmetric. Our locality measure has this desirable property as well. Suppose topic t has members at 3 of S sites. As before, $c_s = 1/S$. When $b_{s,t}$ is [1/3, 1/3, 1/3], $\sigma_t = \sqrt{1/3 - 1/S}$. When $b_{s,t}$ is [1/2, 1/4, 1/4], σ_t increases to $\sqrt{3/8 - 1/S}$. Table 2 shows σ_t as the topic distribution changes from [1/3, 1/3, 1/3] to [4/5, 1/10, 1/10].

3.1.2. Small topics and σ_t

Consider a topic u with a single member. Since there is only one document, only one site, say j, can have it, so

$$\sigma_u = \sqrt{\sum_{s=1}^{S}(b_{s,u} - c_s)^2} = \sqrt{(1 - c_j)^2 + \sum_{s \neq j}^{S} c_s^2}$$

If as before we assume $c_s = 1/S$ then

$$\sigma_u = \sqrt{(1 - 1/S)^2 + (S - 1)(1/S)^2} = \sqrt{(1 - 2/S + 1/S)^2 + (S - 1)(1/S)^2} = \sqrt{1 - 1/S}.$$

If S is reasonably sized (>20), then locality for this topic is both high and fixed. It is not possible to 'de-localize' single member topics. A similar kind of analysis can be made for topics of size m where $m \ll S$. The importance of this analysis is that if there are many small topics, they must be properly accounted for so as not to bias the overall system locality measurement.

3.2. System locality

Given that we have a method to calculate locality on a topic-by-topic basis, the next

Table 2
Topic locality for a topic distributed over 3 of S sites as the distribution changes from uniform to heavily uni-modal

Distribution	σ_t	σ_t, $S = 20$
[1/3, 1/3, 1/3]	$\sqrt{1/3 - 1/S}$	0.532
[1/2, 1/4, 1/4]	$\sqrt{3/8 - 1/S}$	0.570
[3/5, 1/5, 1/5]	$\sqrt{11/25 - 1/S}$	0.624
[2/3, 1/6, 1/6]	$\sqrt{1/2 - 1/S}$	0.671
[4/5, 1/10, 1/10]	$\sqrt{33/50 - 1/S}$	0.781

problem is to identify the proper method to combine a set of topical locality measurements into a single measurement representing the content-locality of the entire system.

In Appendix 1 we show that in general σ_t is bounded between 0 and $\sqrt{2}$ and under equal-sized site assumptions, it is bounded between 0 and 1. For ease of interpretation, we would like the range of the system locality measure to track the range of the topical measure. This suggests the general approach of averaging some or all of the calculated topic locality measurements to get σ. Thus

$$\sigma = \sum_{t \in T'} z_t \sigma_t$$

where

$$\sum_{t \in \mathscr{T}'} z_t = 1, \quad z_t \geq 0$$

The set $\mathscr{T}'$ is the group of topics to be included in the calculation, where $\mathscr{T}' \subseteq \mathscr{T}$ and z_t is a constant that reflects the contribution of each topic to system locality. We consider three variations on setting the values for z_t. The first variation, called Equal, treats the contribution of each topic equally by setting $z_t = 1/|\mathscr{T}'|$. The second method, called Weighted, weights the contribution of a topic according to its overall size, so in this case $z_t = n_t/(\sum_{i \in \mathscr{T}'} n_i)$. Sparse eliminates the contribution of small topics by setting $z_t = 0$ if n_t is less than some threshold and giving topics equal weight if n_t is above the threshold.

The rationale behind the second and third methods is two-fold. As we illustrated in Section 3.1, small topics cannot really exhibit content-uniformity. They are inherently localized. The presence of small topics can give a positive bias to a system locality measurement. The other

	Document Counts				
Topic	S1	S2	S3	S4	n_t
1	0	1	1	0	2
2	4	2	2	2	10
3	2	2	2	1	7
4	0	0	0	1	1

	Site Sizes			
	S1	S2	S3	S4
n_s	6	5	5	4
c_s	0.30	0.25	0.25	0.20

	Content Locality							
t	$\sigma_{1,t}$	$\sigma_{2,t}$	$\sigma_{3,t}$	$\sigma_{4,t}$	σ_t	σ_{Equal}	$\sigma_{Weighted}$	σ_{Sparse}
1	0.090	0.063	0.063	0.040	0.505	0.408	0.185	0.100
2	0.010	0.003	0.003	0	0.122			
3	< 0.001	0.001	0.001	0.003	0.078			
4	0.090	0.063	0.063	0.640	0.925			

Fig. 2. An example calculation of content-locality using a collection with four topics spread over four sites. The table at top left shows the distribution of documents at the four sites. The size of each site is given at top right. At the bottom is locality given individually for each (topic, site) combination, each topic, and for the three methods of calculating system locality. σ_{Sparse} was calculated by eliminating the two smallest topics, 1 and 4.

reason is that the locality measure should accurately reflect the 'strength' of each topic. Weighting each topic equally does not accomplish this.

A concrete example of these three variations is provided in Fig. 2. In this example we show a four site system with four topics. Each site is about the same size, but the topic sizes vary greatly. The last three columns show system locality as measured by the Equal, Weighted and Sparse methods, respectively. Topics 2 and 3 in Fig. 2 have very low topical locality and together make up 85% of the document collection. By weighting all topics equally, the measured locality comes out much higher than the two methods that minimize the contribution of small topics. However, simply throwing out these topics seems too drastic since they are part of the collection. The Weighted method is a reasonable compromise between these two extremes and for this reason is the method of choice.

4. Statistic-based locality

We now turn to an alternative method for measuring content-locality that is based on statistical properties of document collections. In Viles (1996) we showed that when collection statistics differ from that defined by the global corpus, effectiveness can suffer. The method we give here quantifies this difference. The well known *inverse document frequency* (idf) term weighting factor is often calculated as

$$\mathrm{idf}_k = \log \frac{N}{\mathrm{df}_k}$$

for some term k. In a distributed system, each site s has it's own version of statistics derived from the local corpus.

$$\mathrm{idf}_{s,k} = \log \frac{N_s}{\mathrm{df}_{s,k}}$$

If we define a centralized oracle *Cen* that has knowledge of all term statistics at all sites, then we can define the difference in idf for some term k between a site s and the oracle as

$$\delta_{s,k} = \frac{|\mathrm{idf}_{s,k} - \mathrm{idf}_{Cen,k}|}{\log(N)} \tag{2}$$

where we assume term k is found in some document at both sites[1]. The *statistical locality* of term k at site s is $\delta_{s,k}$.

The denominator of Eq. (2) is a normalization factor that scales the quantity between 0 and 1. To obtain the content locality, δ_s, at any site s, we sum the locality measures for every term present in the local collection, C

[1] If k does not exist at s, then we ignore this term in the locality calculation. If k is absent from s, then no document contains it at s. Therefore any query containing k will not match any of these documents on k *even if those documents were located at the Oracle*. So k makes no contribution to the similarity calculation for that document and should be ignored.

$$\delta_s = \frac{1}{K'} \sum_{k \in (C_s \cap C_{\mathrm{Or}})} \delta_{s,k}.$$

where K' is the number of unique terms in $C_s \cap C_{\mathrm{Or}}$. The overall system locality δ is then the average of the locality at each site.

$$\delta = \frac{1}{S} \sum_{s=1}^{S} \delta_s.$$

5. Topic-based locality in the TREC collection

5.1. Data decomposition

To measure topic-based locality we used substantive subsets of the TREC data (Harman, 1995; Voorhees & Harman, 1996). The TREC data comes from multiple sources, consisting of documents from AP Newswire (1988–1990), Wall Street Journal (1987–1992), Computer Select, Federal Register (1988 and 1989), San Jose Mercury News (1991), abstracts from DOE publications and US Patents (1993). Several sources cover multiple years or time periods.

Table 3
Representation of the five topic sets among the 17 document sets of the TREC data. The nature of the TREC experiments means that not all document sets contribute to each topic set. Document counts taken from Callan et al. (1995)

Name	Documents	Topic sets represented				
		1–50	51–100	101–150	151–200	201–250
AP 88	79,919	X	X	X	X	X
AP 89	84,678	X	X	X	X	
AP 90	78,321		X	X		X
DOE	226,087	X	X	X	X	
Fed. Reg. 88	19,860	X	X	X	X	X
Fed. Reg. 89	25,960	X	X	X	X	
Patent	6,711		X	X		X
SJMN 91	90,257		X	X		X
WSJ 87	46,448	X	X	X	X	
WSJ 88	39,904	X	X	X	X	
WSJ 89	12,380	X	X	X	X	
WSJ 90	21,705	X	X	X	X	X
WSJ 91	52,652	X	X	X	X	X
WSJ 92	10,163	X	X	X	X	X
ZIFF 1	75,180	X	X	X	X	
ZIFF 2	56,920	X	X	X	X	X
ZIFF 3	161,021		X	X		X

Table 4
Measurements of topic-based content-locality for the five topic sets of the TREC collection

Topic set	Sites	$\sigma_{\mathrm{Weighted}}$	σ_{Equal}
1–50	13	0.513	0.489
51–100	17	0.389	0.396
101–150	17	0.399	0.401
151–200	13	0.443	0.445
201–250	10	0.409	0.439

There is as yet no generally agreed upon decomposition of the TREC data to do distributed information retrieval experiments, though several have been proposed and used (Walczuch, Fuhr, Pollman, & Sievers, 1994; Callan et al., 1995; Voorhees, 1996; French, Powell, Viles, Emmett, & Prey, 1998) with a general trend of decomposition into more and more sites. One natural way to consider the TREC data as a distributed collection is to make each source and year a site. This is the method that is used in work reported by Callan et al. (1995) and Voorhees, Gupta, and Johnson-Laird (1995) on the 'collection fusion' problem and is the data decomposition we used in the experiments reported here[2]. This set of candidate sites is described in Table 3.

5.2. Topic identification

The major hurdle in the calculation of content-locality is topic identification. The method we use for the TREC corpus is to treat each query as a topic. The documents relevant to that query are considered to be the members of the topic. For the large collections in particular, this leaves a large number of documents that do not belong to an identified topic because they are not relevant to any of the queries provided. This is somewhat unsatisfying, since the disposition of these documents may have an effect on the actual content-locality of a particular collection. However, we can consider these documents as members of unidentified topics which, if known, would have been handled as the known topics were. The identified topics are then considered representative of the universe of possible topics and conclusions drawn from the accompanying results are valid. This kind of assumption has long been assumed in experimental IR work. The possibility of bias in the set of queries is one reason multiple collections are used in IR experimentation.

As we have mentioned, we used the group of topics provided with the TREC collection and the set of accompanying relevant documents to identify the members of each topic. However, because of the nature of the TREC experiments, not all of the subcollections identified in Table 3 have relevance judgements for all of the five, 50 member TREC topic sets (numbered 1–50, ..., 201–250) we used in this study. For example, the topic members for topic set 201–250 have been identified for only 10 sites. When we calculate locality for any particular topic set, we can use only the subcollections for which the topic members have been identified.

[2] *Collection fusion* is the process of merging results from searches performed on different collections.

5.3. Results

Because we have five sets of topics, we generated five measurements of topic-based system locality. These measurements appear in Table 4. Since they are single measurements, it is hard to assess whether the difference between topic sets is significant. However, the absolute differences are small.

In Table 4 we also give the unweighted locality measure, σ_{Equal}. These match closely with $\sigma_{\mathrm{Weighted}}$, indicating that there is relatively little 'small topic effect' in these measurements. This observation is also supported by a scatterplot plotting locality against topic size (Fig. 3). In this plot, we show five sets of TREC topics using different symbols, something which TREC-initiated readers may find helpful. Otherwise, the plot can be interpreted as a simple scatterplot and the difference between symbols ignored. Regardless, there appears to be little correlation between locality and topic size.

6. Statistic-based locality in NCSTRL

The Networked Computer Science Technical Report Library (http://www.ncstrl.org) is a distributed collection of technical reports from over 100 academic sites doing research in Computer Science. It has been operational since July of 1995 and currently services thousands of queries per day. Here we measure the content locality of this operational, distributed document collection using the statistically based measure, δ.

The data we used for this analysis was obtained at two different times, late July of 1995

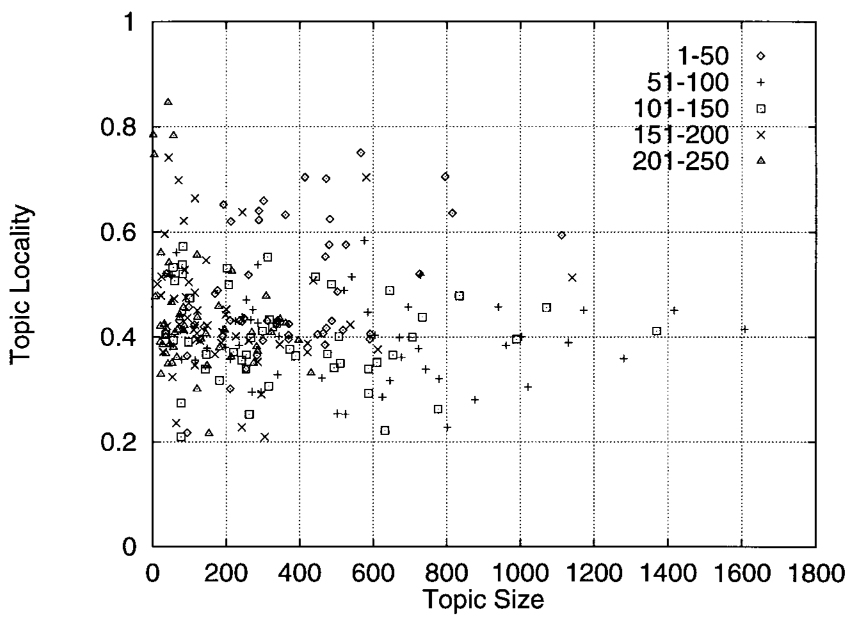


Fig. 3. Topic size versus locality in the 250 TREC topics.

Table 6
Statistic-based content-locality of the NCSTRL collection at two times, July 1995 and February 1998. At top is the locality of the 18 sites who were members of NCSTRL at both times. Summary locality measurements for February 1998 are given for both the 18 'original' NCSTRL sites as well as for the entire NCSTRL archive. The sites are arranged in order of increasing locality of the July 1995 measurement

Site	July 1995		February 1998	
	size	locality (δ_s)	size	locality (δ_s)
MIT	2342	0.070	2348	0.094
Cornell	1412	0.089	1532	0.109
Stanford	758	0.104	1230	0.114
Cal-Irvine	477	0.114	564	0.124
Wisconsin	521	0.118	643	0.130
Virginia Tech	420	0.120	480	0.132
Cal-Berkeley	968	0.121	1107	0.127
Hong Kong	30	0.125	30	0.124
Virginia	309	0.131	371	0.115
Princeton	188	0.134	274	0.161
Auburn	86	0.134	86	0.137
Maryland	293	0.136	595	0.143
Chicago	137	0.137	276	0.137
SUNY Buffalo	120	0.146	188	0.141
Old Dominion	93	0.147	183	0.152
Boston U.	48	0.156	107	0.185
UNC-Chapel Hill	95	0.159	155	0.170
Iowa State	100	0.164	123	0.180

	System locality (δ)		
	July 1995	February 1998 ($n=18$)	February 1998 ($n=73$)
Average	0.128	0.138	0.162
Standard deviation	0.024	0.024	0.050
CV (percent)	18.8	17.4	30.6

from a beta-version of NCSTRL and early February 1998. Gross characteristics of the collection at these two times are given in Table 5.

Table 5
Characteristics of the NCSTRL collection at two times

	July 1995	February 1998
Number of sites	29	102
Number of 'large' sites (≥30 docs)	18	73
Number of docs	8450	21,357
Number of docs at large sites	8397	21,158

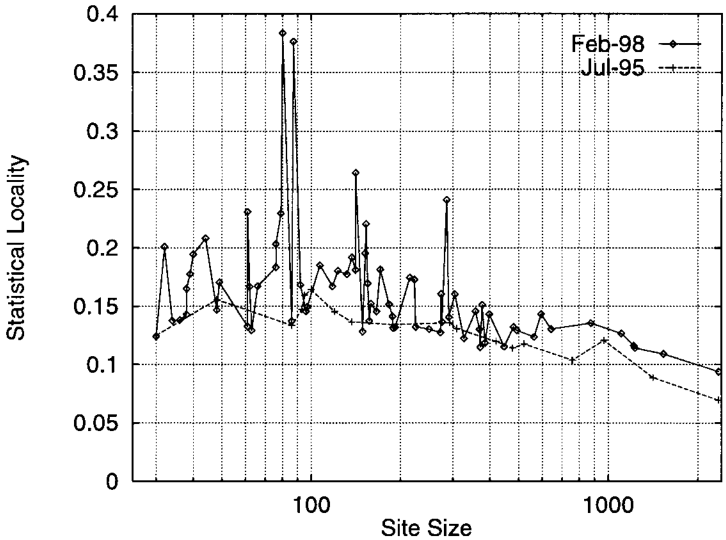


Fig. 4. Site size versus statistical locality in NCSTRL at two time periods, July 1995 (18 sites) and February 1998 (73 sites).

Each document in NCSTRL is a bibliographic record that includes at a minimum, author, date and title. A large percentage of the records contain abstracts and on-line full text in a variety of formats (OCR, postscript, page images). We make the normal IR assumption that the available terms in a text adequately express the topical content of the document it represents (Salton & McGill, 1983). For our locality measurements, we considered only the text contained in the title and abstract. This text was preprocessed by removing common words and stripping words to their stems. Using the δ measure, we then compared the statistical signature of each site against a central collection composed of the entire bibliographic collection.

Results of this operation are given in Table 6 along with the collection size of each site. There is a rough, inverse correlation between the size of the site and the locality of that site, which is depicted graphically in Fig. 4. This is to be expected. The older, well-established departments have larger collections and tend to be more representative of the entire collection than smaller departments which tend to focus on a small number of selected research areas. For example, the SUNY-Buffalo archive is heavily theory oriented and the Iowa State archive has emphasis in languages and search algorithms. Fig. 4 also shows that the NCSTRL collection has grown more content-localized over time and that the site-to-site variation of measured locality has also increased.

As we have noted, the content-locality for the 18 ‘original’ NCSTRL sites increased over time. In July 95, each site made up a larger proportion of the NCSTRL DL than it does now and thus we would expect it to be more representative of the entire DL. At the later time, each site is less representative, thus content-locality should be higher. This is exactly the effect we observed.

Examination of Table 6 also yields some insight into the challenges of fielding operational distributed digital libraries (Lagoze, Fielding, & Payette, 1998). For example, the Computer Science Departments at several sites show little if any growth in their archives though in reality they continue to produce technical reports. This is related to software version difficulties and technical support considerations rather than actual report production. Though some of the original sites have not contributed any new documents to the DL, their content locality has changed. This is to be expected, because content-locality is measured with respect to the entire DL and the entire DL has changed considerably in between the two snapshots.

We also note that the average measured locality of NCSTRL increased from 0.128 to 0.162 over the 30 month time period and the coefficient of variation (CV) increased as well, from 18.8 to 30.6%. As the DL grows, it is becoming more content-localized, not less.

7. Discussion

Operational distributed document collections are only now starting to be deployed. The usefulness of content-locality monitoring is still undetermined. The major motivation in the context of this work is determining whether or not member sites in the distributed document collection need to communicate statistical information about their local collections to other member sites. If locality is low, then no communication is needed. If locality is high, then communication is needed to maintain good retrieval effectiveness.

Of course, this begs the question about what exactly is 'low' and 'high'. Clearly, if $\sigma = 0.01$ then locality is low and if $\sigma = 0.98$ locality is high. However, if $\sigma = 0.40$, then in what situation are we? There are two questions to answer. First, how 'skewed' is a topic that shows locality of 0.40? Second, does $\sigma = 0.40$ mean the system will show reduced search effectiveness?

To address the first question, we selected three topics from the TREC topics showing 'low', 'average' and 'high' locality relative to the observed locality of the entire 250 topics. The text

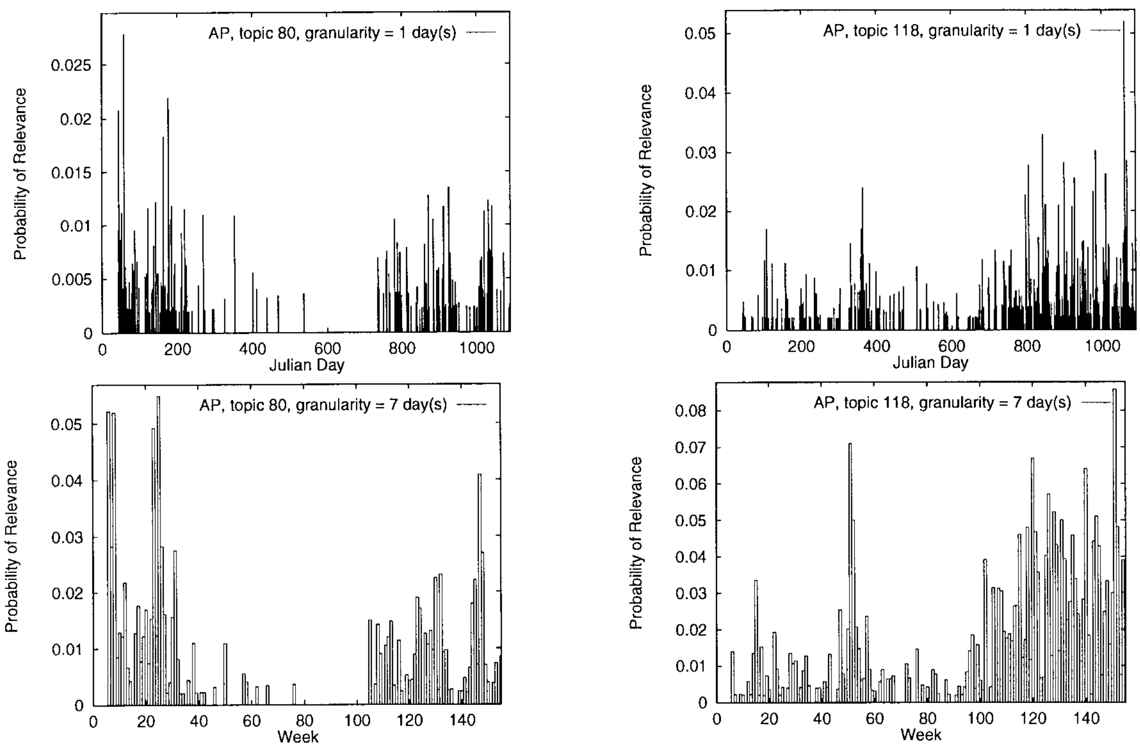
Topic	Text Excerpt
161	*Document will provide information on the problems and actions associated with what is known as acid rain.*
152	*Accusations of Cheating by Contractors on U.S. Defense Projects. Document will refer to an alleged illegality committed by any entity seeking a contract on behalf of the U.S. Military Forces.*
37	*Document identifies software products which adhere to IBM's SAA standards. To be relevant, a document must identify a piece of software which is considered a Systems Application Architectural (SAA) component or one which conforms to SAA.*

	Sub Collection														
Topic	AP88	AP89	DOE	FR88	FR89	WJ87	WJ88	WJ89	WJ90	WJ91	WJ92	Z1	Z2	n_t	locality
161	18	25	25	2	1	6	4	3	5	8	1	0	1	305	0.21
152	30	28	0	0	0	7	16	3	5	5	1	2	3	535	0.42
37	0	1	0	0	0	0	0	0	1	1	0	25	72	566	0.75

Fig. 5. Three topics from the TREC collection. Excerpts of text from the topics are at top and the percent distribution of each topic at each site is at the bottom. Topics were chosen to reflect 'low' (topic 161), 'average' (topic 152) and 'high' (topic 37) locality.

of these three topics appears at the top of Fig. 5 and the distribution of the topic among the 13 TREC 'sites' is given at the bottom of Fig. 5. The topic distribution gives some useful intuition into what a value of σ means operationally. The distribution of the 'low' locality topic 161 is concentrated at three main sites, but is represented at all sites except one. Topic 152, the medium locality topic, is also concentrated at three sites, but is represented at fewer sites than topic 161, so locality is higher here. At the 'high' locality end, topic 37 is represented almost entirely at two sites with only token representation at three other sites. Locality is highest for this topic.

Regarding the second question, it is reasonable to conclude that the NCSTRL archive taken as a whole is a skewed collection, the wide variation in individual site localities lends credence to this conclusion. However, we cannot definitively say that, of 0.138 means that sites need to



Topic	Text Excerpt
80	*Document will identify something about the platform of a 1988 presidential candidate.*
118	*Document will provide background information on international terrorist groups or individuals, or detail the activities of such groups or individuals.*

Fig. 6. An example of content-locality in time. This figure shows the temporal distribution of relevant documents in the AP Newswire (1988–1990) for two TREC topics at granularities of 1 and 7 days. One topic asks about presidential politics, the other asks about terrorist activities.

communicate with each other in order to maximize effectiveness. What is needed is systematic, concurrent monitoring of both content-locality and user satisfaction in operational systems. With the deployment of working systems like NCSTRL, the ability to do such work is only now becoming possible.

7.1. Temporal locality

In addition to content-locality in space, the possibility of locality in time exists as well. In temporally ordered document collections like news archives, there is a form of topical clustering in time that naturally arises in topics that are related to events (natural disasters, elections), days of the week (church services, Friday night football games) and seasons (snow, fall foliage, summer beach traffic).

Most work focusing on ad-hoc queries assumes that the temporal distribution of relevant documents is uniform. That is, the probability of relevance is independent of when the document was created. Intuitively, this appears to be an invalid assumption for at least the kinds of queries outlined above. In such cases, knowledge of the time-based probability distribution of relevance might aid considerably in focusing retrieval.

Focused work on event detection and tracking using transcriptions of news broadcasts has shown that the explicit use of temporal features can help with retrieval. For example, Allan, Papka, and Lavrenko (1998) explicitly factor in time when determining whether a news story is part of a previously detected event or describes a new event while Yang, Pierce, and Carbonell (1998) use temporal proximity as a feature in event-focused document clustering.

Fig. 6 provides evidence to support this intuition. Depicted is the temporal distribution of relevance in a 3 year period of the AP Newswire for two TREC topics. In both cases, the distribution of relevance is decidedly non-uniform.

8. Summary

The notion of content-locality in distributed document collections is new. The work we describe in this paper is an effort to more fully understand the underlying phenomenon. Through the introduction of two possible measurement methodologies, we have made considerable progress in this regard. Using these methods, we measured the topic locality of the TREC collection and statistical locality of an operational distributed collection, NCSTRL.

The fidelity of the topic-based locality measurement rests entirely on the accuracy of topic determination. If good topic descriptions (e.g. subject descriptions) are available, then locality is easy to calculate. Content-locality is only one of many reasons that topic identification is worth knowing. Others include improved efficiency, faster retrieval and more effective browsing. Comparison of content-locality derived from different systems should not be done lightly. The method of topic identification is the key to enabling a reasonable comparison. To the extent that topic identification differs between systems, direct comparison may become increasingly meaningless.

The statistic-based locality measure has potential of being implemented operationally because it is wholly automatic and easy to calculate. No topic determination is needed. Scaling problems need to be overcome and methods to account for or remove rare term and small collection artifacts need to be devised.

Since temporal topic locality exists, there are a number of follow-up questions to pursue. How do we measure it? Can we gain insight from other phenomenon that exhibit locality, e.g. memory reference patterns? Is there a detectable attribute in the query that might indicate a temporal component to relevance? If so, then knowledge of the 'highly relevant' region(s) might significantly improve retrieval effectiveness for these queries through focused search in the region(s). What is the relationship between 'topic' and 'event' (Allan et al., 1998)? Are there distinct types of temporal patterns associated with topics, e.g. uni-modal, multimodal, periodic? Again, such knowledge could further focus search efforts.

Open questions remain about how best to take advantage of content locality in distributed digital libraries. Locality is highly context sensitive, term distributions may show high locality in one distributed archive and low locality in another. What to do with, for example, statistical information then becomes context sensitive as well. Sometimes sites may need to share information to achieve high effectiveness on a search, while other times such sharing may not be needed.

Locality in time and space is not a new concept, being integral in a wide variety of areas including analysis of memory reference patterns (Madison & Batson, 1976; Weikle, McKee, & Wulf, 1998), file caching in networked file systems (Satyanarayanan, 1989) and caching in various distributed computer systems, World Wide Web servers and browsers being the most obvious current example. Careful consideration of this literature may provide deeper insight into the nature of content-locality and methods to measure it.

Acknowledgements

We thank the reviewers for their helpful comments. This work was supported in part by NASA Goddard Space Flight Center under GSRP Fellowship NGT-51018 and by contract No. N66001-97-C-8542 awarded by DARPA.

Appendix A. Notes on σ

A.1. Bounds on σ_t

From the definitions of c_s and $b_{s,t}$ Table 1 we get

$$\forall s\colon\ 0 \le c_s \le 1 \quad \text{and} \quad \sum_{s=1}^{S} c_s = 1$$

and

$$\forall s,t\colon\ 0 \le b_{s,t} \le 1 \quad \text{and} \quad \sum_{s=1}^{S} b_{s,t} = 1$$

These constraints mean that we can put bounds on σ_t, specifically, $0 \leq \sigma_t < \sqrt{2}$. The only time $\sigma_t = 0$ is when $c_s = b_{s,t}$, $\forall s,t$. To see that $\sigma_t < \sqrt{2}$, consider the following:

$$\sigma_t = \sqrt{\sum_{s=1}^{S} \sigma_{s,t}} = \sqrt{\sum_{s=1}^{S} (b_{s,t} - c_s)^2}$$

$$= \sqrt{\sum_{s=1}^{S} b_{s,t}^2 - 2\sum_{s=1}^{S} c_s b_{s,t} + \sum_{s=1}^{S} c_s^2} \leq \sqrt{1 - 2\sum_{s=1}^{S} c_s b_{,st} + 1} \leq \sqrt{2 - 2\sum_{s=1}^{S} c_s b_{s,t}} < \sqrt{2}$$

The situations when σ_t is close to $\sqrt{2}$ are when a topic is completely located at a site that is very small relative to other sites. A simple example is a 2 site system where $b_{s,t} = (0, 1)$ and $c_s = (0.9, 0.1)$. In this case we get

$$\sigma_t = \sqrt{\sum_{s=1}^{S} (b_{s,t} - c_s)^2} = \sqrt{(0 - 0.9)^2 + (1 - 0.1)^2} = \sqrt{1.62}$$

However, if sites are approximately the same proportion, namely $1/S$, then we can derive a tighter bound than $\sqrt{2}$:

$$\sigma_t = \sqrt{\sum_{s=1}^{S} (b_{s,t} - 1/S)^2} = \sqrt{\sum_{s=1}^{S} b_{s,t}^2 - \frac{2}{S}\sum_{s=1}^{S} b_{s,t} + \sum_{s=1}^{S}\left(\frac{1}{S}\right)^2} = \sqrt{\sum_{s=1}^{S} b_{s,t}^2 - \frac{1}{S}} < 1.$$

A.2. Properties of σ_t

To appreciate the behavior of σ_t, we first consider the set of systems where a topic is uniformly distributed over $k = 1, 2, \ldots, S$ sites. The locality measure should decrease monotonically as k increases. For ease of analysis, assume that the size of each site c_s is constant i.e. $\forall s$, $c_s = 1/S$. When $k = 1$,

$$\sigma_t = \sqrt{\sum_{s=1}^{S} (b_{s,t} - c_s)^2} = \sqrt{(1 - 1/S)^2 + (S - 1)(1/S)^2} = \sqrt{1 - 2/S + 1/S^2 + (S - 1)(1/S^2)}$$

$$= \sqrt{1 - 1/S}.$$

When $k = 2$,

$$\sigma_t = \sqrt{2(1/2 - 1/S)^2 + (S - 2)(1/S^2)} = \sqrt{2(1/4 - 1/S + 1/S^2) + (S - 2)(1/S^2)}$$

$$= \sqrt{1/2 - 1/S},$$

and for arbitrary k,

$$\sigma_t = \sqrt{k(1/k - 1/S)^2 + (S-k)(1/S^2)} = \sqrt{k(1/k^2 - (2/k)(1/S) + 1/S^2) + (S-k)(1/S^2)}$$
$$= \sqrt{1/k - 1/S}.$$

References

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In: *Proceedings of the 21st Conference on Research and Development in Information Retrieval*. Melbourne, Australia.

Callan, J. P., Lu, Z., & Croft, W. B. (1995). Searching distributed collections with inference networks. In: *Proceedings of the 18th International Conference on Research and Development in Information Retrieval* (pp. 21–29). Seattle, WA.

Cutting, D., Karger, D., Pedersen, J., & Tukey, J. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In: *Proceedings of the 15th International Conference on Research and Development in Information Retrieval* (pp. 318–329). Copenhagen, Denmark.

Davis, J. R. (1995). Creating a networked computer science technical report library. In *D-Lib Magazine* http://www.dlib.org/dlib/september95/.

French, J. C., Powell, A., Viles, C. L., Emmitt, T., & Prey, K. (1998). Evaluating database selection techniques: a testbed and experiment. In: *Proceedings of the 21st Conference on Research and Development in Information Retrieval*. Melbourne, Australia.

French, J. C., & Viles, C. L. (1996). Ensuring retrieval effectiveness in distributed digital libraries. *Journal of Visual Communication and Image Representation*, *7* (1), 61–73.

Gravano, L., Chang, C. H., Paepcke, A., & Garcia-Molina, H. (1997). STARTS: Stanford proposal for Internet meta-searching. In *Proceedings of the 1997 International Conference on the Management of Data (SIGMOD '97)*.

Gravano, L., & Garcia-Molina, H. (1997). Merging ranks from heterogeneous information sources. In: *Proceedings of the 23rd International Conference on Very Large Databases (VLDB'97)*. Zurich, Switzerland.

Harman, D. (1995). Overview of the 4th Text Retrieval Conference (TREC-4). In: *Proceedings of the 4th Text Retrieval Conference (TREC-4)*. Gaithersburg, MD.

Kellogg, R. B., & Subhas, M. (1996). Text to hypertext: can clustering solve the problem in digital libraries? In: *Proceedings of the First ACM International Conference on Digital Libraries* (pp. 144–150). Bethesda, MD.

Lagoze, C., Fielding, D., & Payette, S. (1998). Making global digital libraries work: collection services, connectivity regions and collection views. In: *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 134–143). Pittsburgh, PA.

Madison, A. W., & Batson, A. P. (1976). Characteristics of program localities. *Communications of the ACM*, *19* (5), 285–294.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Satyanarayanan, M. (1989). Distributed file systems. In S. J. Mullender, *Distributed systems*. ACM Press.

Viles, C. L. (1996). *Maintaining retrieval effectiveness in distributed, dynamic information retrieval systems*. Ph.D. thesis, University of Virginia.

Viles, C. L., & French, J. C. (1995). Dissemination of collection wide information in a distributed information retrieval system. In: *Proceedings of the 18th International Conference on Research and Development in Information Retrieval* (pp. 12–20). Seattle, WA.

Voorhees, E. (1996). The TREC-5 database merging track. In: *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*. Gaithersburg, MD.

Voorhees, E., Gupta, N. K., & Johnson-Laird, B. (1995). Learning collection fusion strategies. In: *Proceedings of the 18th International Conference on Research and Development in Information Retrieval* (pp. 172–179). Seattle, WA.

Voorhees, E., Harman, D. (1996). Overview of the 5th Text Retrieval Conference (TREC-5). In: *Proceedings of the 5th Text Retrieval Conference (TREC-5)* (pp. 1–28). Gaithersburg, MD.

Walczuch, N., Fuhr, N., Pollman, M., & Sievers, B. (1994). Routing and ad-hoc retrieval with the TREC-3 collection in a loosely federated environment. In: *The 3rd Text REtrieval Conference (TREC-3)* (pp. 135–144). Gaithersburg, MD.

Weikle, D. A. B., McKee, S. A., & Wulf, W. A. (1998). Caches as filters: a new approach to cache analysis. In: *6th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*. Montreal, Canada.

Yang, Y., Pierce, T., & Carbonell, J. (1998). A study on retrospective and on-line event detection. In: *Proceedings of the 21st Conference on Research and Development in Information Retrieval*. Melbourne, Australia.